Large-scale implementation of pooled RNA extraction and RT-PCR for SARS-CoV-2 detection - Supplementary note

Here we will elaborate on the efficiency of matrix pooling. The first stage of testing is sufficient for identifying the positive samples in the matrix as long as they are either all on the same row or all on the same column. This is the case, for example, if there is only one positive sample in the matrix. As will be shown below, the probability that no retesting will be needed is $2nq^{n^2-n} - (2n-1)q^{n^2} - n^2pq^{n^2-1}$, where $q = 1 - p$, and so for $p = 1\%$ and $n = 5$, less than 2% of the pools will require retesting. Thus, for $p = 1\%$, using a 5x5 matrix leads to an almost 2.5-fold increase in throughput (or 40% of tests), as most matrices require about 10 tests instead of 25, and an 8x8 matrix gives already gives an efficiency of almost 4, at the price of an increased probability of retesting of around 11%.

The above formula for the probability of no retesting can be derived as follows: denote by A the event that at most 1 row tested positive and by B the event that at most 1 column tested positive. These events have the same probability, which is $q^{n^2} + nq^{n^2-n}(1 - q^n)$. The event of no retesting is the union of A and B. It is easier to calculate the probability of their intersection which is exactly the event that there is at most one positive sample in the matrix, hence its probability is $q^{n^2} + n^2pq^{n^2-1}$. The probability of the intersection of A and B is the sum of their probabilities minus the probability of their union, which gives the aforementioned formula.

The efficiency of the matrix algorithm is:

$$\left[\frac{2}{n} + (1 - q^n)\left(1 - q^{n^2-n}\right) - q^n(1 - q^{n-1})\left(1 - q^{(n-1)^2} + pq^{n^2-2n}\right)\right]^{-1}.$$

A derivation of this formula appears in the next paragraph. Theoretically, for very low values of p, the optimal pool size n is roughly $p^{-\frac{2}{3}}$, resulting in an efficiency of about $\frac{1}{3}p^{-\frac{2}{3}}$, which is asymptotically higher than Dorfman pooling (but asymptotically lower than the information bound, which is roughly $\frac{1}{p(\frac{1}{p})}$). However, in practice, the pool size is limited to values which allow safely identifying a single positive sample in the pool. When keeping the pool size fixed, as the probability p becomes smaller, Dorfman pooling becomes about twice as efficient as matrix pooling, because it tests every sample in only one pool instead of two, and most pools test negative.

A derivation for the efficiency of the matrix algorithm is as follows. The efficiency is the number of samples, n^2, divided by the expected number of tests conducted on the matrix. The number of tests in the first stage is always $2n$. The expected number of tests in the second state is n^2 times the probability that a fixed sample will need to be retested. We will calculate this probability for the sample lying the first row and first column. Denote by R_i (resp. C_i) the events that the ith row (resp. ith column) tested positive. Let R (resp. C) be the event that the first row (resp. column) and at least one more row (resp. column) tested positive. We wish to find the probability of the intersection of R and C. Since R is the event that there is a positive sample in both the first row and in the rest of the matrix, its probability is $(1-q^n)\left(1-q^{n^2-n}\right)$. In order to compute the probability of $R \cap C$, it is enough to find the probability of the difference $R \backslash C$, which can be decomposed as the disjoint union of the events D_1 and D_2, where $D_1 = R \backslash C_1$ and D_2 is the intersection of R with the event that only the first column tested positive out of all columns. D_1 is the event that the first column is negative but the first row and at least one more row is positive, hence its probability is $q^n(1-q^{n-1})(1-q^{(n-1)^2})$. The probability of D_2 can similarly be seen to be $p(1-q^{n-1}) \cdot q^{n^2-n}$. The probability of $R \backslash C$ is the sum of the probabilities of D_1 and D_2, and the probability of the intersection of R and C is the difference in probabilities of R and $R \backslash C$. Multiplying by n^2 gives us the expected number of retests needed, from which the matrix efficiency formula immediately follows.

Many variants of the matrix pooling described above may be considered for the purpose of lowering the probability of a second testing stage. One such approach may be using d-disjunct testing matrices (see [1]) as a first stage. This approach allows accurate (and efficient) identification of the positive samples from the test results if there are at most d positive samples. By comparison, the matrix algorithm described above will generally require a second stage when the matrix contains two positive samples. Another possible variant uses a d-dimensional matrix of side length n instead of the 2-dimensional matrix. In such a setting, at first stage, each generalized row is tested, totalling dn^{d-1} tests. For example, if $d = 3$, we test each horizontal row, vertical row and depth row. A sample in the cube is suspected positive if all its generalized rows tested positive. If a sample suspected positive has the additional property that at least one of its generalized rows contains no other suspected positive sample, then we can deduce it is positive. If all suspected positive samples can be deduced positive this way, then no further testing is needed. Note that, in addition to a possible lower probability of retesting, d-dimensional matrix pooling offers a theoretically higher asymptotic efficiency of order $p^{-\frac{d}{d+1}}$ for very low a priori probabilities. However, such efficiency is attained at an optimal pool size of the same order, i.e. roughly $n = p^{-\frac{d}{d+1}}$, and so might not be practical for clinical purposes.

1. Du, D., Hwang, F. K. & Hwang, F. *Combinatorial Group Testing and Its Applications*. (World Scientific, 2000).